[[what's abstract ??]]
Word counts:
Introduction 508
BMR 797
Rabit 729
Rewiring 652
Validation 470
Conclusion 369

A large scale integrative resource from ENCODE for cancer research

Introduction

A small fraction of mutations associated with cancer have been well characterized, particularly those in coding regions of key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – especially those discovered over the course of recent whole genome cancer genomics initiatives – lie within non-coding regions \cite{25261935}. Whether these mutations have substantial functional impact on cancer progression remains an open question \cite{26781813}.

Several recent studies have begun to address this question by incorporating limited functional genomics data for variant interpretation \{cite 25261935, 27064257, 27807102}. For example, *Hoadley et al.* integrated five genome-wide platforms and one proteomic platform to uniformly classify various tumor types \{cite 25109877}. *Torchia et al.* integrated various genomic and epigenetic signals to identify promising therapeutic targets in rhabdoid tumors \cite{27960086}. *Lawrence et al.* incorporated large-scale genomics profiles to identify cancer drivers \{cite 23770567}. However, there is no systematic integration of thousands of functional genomic data sets from a broad spectrum of advanced assays to interpret the cancer genome.

The rich functional assays and annotation resources developed by the ENCODE Consortium allows us to characterize these non-coding regions at a great depth \cite{22955616}. Data from ENCODE is particularly suited for cancer research as around eighty percent of the ENCODE cell lines are associated with cancerous tissues (see supplements). In the initial release of the ENCODE annotation, this was predominantly accomplished by using RNA-seq and ChIP-seq assays on a limited number of cell types \cite{22955616}. The new release of ENCODE took two new directions. First, it considerably broadened the number of cell types with RNA-seq, ChIP-seq, and DNase-seq assays, hence the main ENCODE encyclopedia aims to utilize these to provide a general, unified annotation resource applicable across many cell types. Secondly, ENCODE also expanded the number of advanced assays such as STARR-seq, Hi-C, ChIA-PET, eCLIP and RAMPAGE on several top-tier cell lines. Many of which are cancer-associated, including the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervical (HeLa-S3) cancers (Figure 1A). In addition, another data enriched top-tier cell line H1-hESC is from a human stem cell. It has been thought for decades that at least a subpopulation of the tumor cells have the ability to self-renew, differentiate, and regenerate, similar to what is conceptualized in normal stem cells \cite{24333726}. Hence, H1-hESC can serve as a valuable comparison to cancer cells to check the degree to which their oncogenic transformation is in a more differentiated or undifferentiated direction \cite{24333726}.

Here, we endeavor to collect the data catalog to provide deep annotations of cancer genomes. We performed large-scale integration to construct an in-depth cancer related companion resource to the general encyclopedia. We complied these resources as the "companion *ENCODE* encyclopedia resource for *C*ancer" (or "EN-CODEC" for short) to interpret cancer-related genomic data, such as mutational and transcriptional profiles. [[JZ2MG: EN-CODEC or quoted?]]

Formatted: Highlight

Multi-level data integration improves variant recurrence analysis in cancer

One of the most powerful ways of identifying key elements in cancer is through recurrence analysis to discover regions that mutate more than expected. Hence, we first attempted to construct an accurate background mutation rate (BMR) model in a wide range of cancer types. However, this is a challenging problem since the somatic mutation process can be influenced by numerous confounding factors (in the form of both external genomic factors and local sequence context factors), which without appropriate correction can result in many false positives or negatives \{cite 23770567}. Here, we tackled these problems by removing effects of confounding factors in a cancer-specific manner. Specifically, we separated the whole genome into bins (1Mb) and calculated mutation counts per bin under each local context category. For each category, we used a negative binomial regression of the mutation counts against 475 features from xx cell lines, including replication timing, chromatin accessibility, Hi-C, and expression profiles for BMR prediction. In contrast to methods that use unmatched data \{cite 23770567}, our approach automatically selects the most relevant features, thereby providing noticeable improvements in BMR estimation (Fig 2A). Notably the combination of many different genomic significantly improves the estimation accuracy in multiple cancer types (Fig 2 B). Also, it is worth mentioning that due to the correlated nature of these genomic features, some cancers without features from apparently matched cell types can still automatically learn from related cell types and achieve a decent BMR precision. Hence, our analyses could be easily extended to other cancer types. [[JZ2MG: checking number of cell lines right now]]

A second step to utilize the ENCODE data in the recurrence analysis in cancer is to maximize the statistical power of burden tests. In terms of an individual test, focusing on shorter core regions with true functional impact would significantly improve the computation power. Hence, we first trimmed the conventional annotations, such as enhancers, to the key regions by looking into shapes of various signal tracks (see supplements). Furthermore, different from traditional analysis where comprehensive annotations are usually beneficial, testing every nucleotide in the genome will subject to huge penalty from multiple testing corrections and significantly reduce statistical power (see supplementary file). Therefore, we tried to focus on a minimum number of high confident annotations to look for burdened regions. Particularly for enhancers, we started from searching for regions supported by multiple evidence. We first proposed a machine learning algorithm CASPER to combine shapes of signal tracks from DNase-seq and a battery of 5 to 10 histone modification marks. We then assembled the CASPER predictions with peaks called from STARR-seq experiments, which directly read out candidate enhancers in the genome. Such an integrative approach enables accurate enhancers definitions (see supplement). We also reconciled these enhancers with the main encyclopedia annotations by reporting the overlapped ones and providing new IDs to the novel ones. [[JZ2MG: should we mention the enhancer number here? I prefer not…]]

[[JZ2MG: extended gene section, we can claim that we are increasing power, but also we can claim the following para is to increase the functional interpretbility. Which do you prefer]]

A final aspect to increase the power is to link the compact noncoding regulatory elements to the protein coding genes to form an extended gene region as a whole test unit. A natural consequence of this is, analogous to the exon regions within genes, a set of discrete regions that potentially affect gene expressions. Such unified annotation enables a joint evaluation of the mutational signals from distributed yet biologically relevant genomic regions. Traditional methods have to solely rely on computational correlation due to the lack of data, resulting in problematic extended gene definition. Here we use direct experimental evidence and physical interaction from the Hi-C and ChIA-PET, combined with a machine learning algorithm that takes into consideration of the wide variety of histone modification marks and expressions to achieve accurate enhancer target gene linkages. Finally, the conserved enhancer-target linkages, refined promoters, and RNA-binding sites from eCLIP experiments within genes constitute a so-called extended gene

Deleted: like

Deleted:

Formatted: Highlight

Deleted: annotation

Deleted: burden

Deleted: focus on compact annotation set with high confidence. In fact,

Deleted: large

Deleted: , which could

Deleted: appreciably dilute the mutational signal and

Deleted: provide

Deleted: defining enhancers

Deleted: used

Deleted: genomic

Deleted: in combination with DNase-seq

Deleted: These were used as input into CASPER, a machine learning predictor that we developed to integrate the shapes of these various signals.

Deleted: these

Formatted: Highlight

Deleted: In addition, it is also important to increase power by confining the mutationeach individual mutational [[check]] burden tests on core annotations with shorter length but higher functional impact (see supplement). Hence, we refined the above-mentioned enhancers through the CASPER algorithm by trimming down candidate regions to a smaller size atbased on the centershape of hte histone marksmark peak.

Deleted: [[JZ2MG: we want to be clear otherwise experts like Shirley will keep on asking what did you do exactly? But potentially by only selecting the center of histone mark is dangerous. I have mixed feeling of the last sentence]][[let's disc]]

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Deleted: refined

Formatted: Not Highlight

Deleted: [[JZ2MG: do you understand what does “refined” mean here? I have a feeling that reviewers will not, but haven’t got a better name]], high confidence

Deleted: (the so-called extended gene region)

Deleted: to

Deleted: ly

Deleted: pick up

Deleted: account

neighborhood (Fig1 C). Such joint test scheme also results in much more interpretable burdened regions as they are often associated with well-known oncogenic genes.

We demonstrate that our multi-level integration scheme can effectively remove false positives and discover meaningful regions with higher-than-expected mutation counts (Fig 2C). For example, in the context of chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes, such as TP53 and ATM, which has been reported from previous coding region analysis. It also discovered genes that are missed by the exclusive analysis of coding regions, such as BCL6. Note that BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene neighborhood could be used as an annotation set for recurrence analysis.

Integrating regulatory networks and tumor expression profiles identifies key regulators in cancer

ENCODE annotation also provides detailed regulatory networks instantiated from experimental assays suitable for cancer research. Specifically, for the TF network we first built distal and proximal TF regulatory networks by linking TF to genes, either directly by TF-gene interactions by promoters or indirectly via TF-enhancer-gene interactions in each cell type (Fig1 B). We then pruned these networks to include only the strongest edges using another signal shape algorithm \{cite 22039215}. In addition, we merged our cell-type-specific networks to get a generalized network for pan-cancer analysis. Similar, we also defined an analogous RBP network in a simpler format. Compared to imputed networks from motif analysis, our ENCODE TF and RBP regulatory networks were built upon actual ChIP-seq and eCLIP expriments, which provide much more accurate regulatory interactions between functional elements.

Deleted: called TIP

Deleted: , which is a valuable resource for interpreting the wealth of cancer gene expression from cancer tissue

The integrated networks are useful for interpreting the oncogenic changes evident in cancer gene expression data from tumor samples. In particular, using a machine learning method, we integrated 8,202 tumor expression profiles from TCGA to systematically search for the TFs and RBPs that most strongly drive tumor-specific expression patterns. For each patient, our method tests to the degree a regulators' regulation potentials are sufficiently correlated with their targets' tumor-to-normal expression changes. We then calculated the percentage of patients with these relationships in each cancer type and presented the overall trends for key TFs and RBPs in Fig. 3A.

We find that the target genes of MYC are significantly up-regulated in numerous cancers, which is consistent with its well-known role as an oncogenic TF and a transcription activator \cite{22464321}. We further validate MYC's regulatory effect through knock down experiments (Fig 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown (Fig 3A). After confirming the importance of MYC, we can use the regulatory network to understand how MYC works with other TFs. We first looked at all triplets involving MYC by requiring that a second TF both interacts and shares a common target with MYC. In all cancer types, we found that MYC's expression levels are positively correlated with the expressions of most of its targets, while the second TF shows only a limited influence as determined from partial correlations. We then investigated the exact structure of such regulatory relationships. The most common triplet interaction mode is a well-understood feed-forward loop (FFL) structure in which MYC regulates both the common target and the second TF. Most of these FFLs involve well-known MYC partners such as Max and Mxl1. However, we also discovered that many involve another factor called NRF1. Upon further study, we found that that the MYC-NRF1 FFL relationships were mostly coherent ("amplifying"). We further studied these FFLs by forming these triplets into a logical gate, in which the two TFs act as inputs and the target gene expression represents the output \{cite 25884877}. We can show that the predominant number of these gates follow either OR or MYC-always-dominant logic. Thus, the ENCODE regulatory network not only helps find key regulators, but also demonstrates how they work in combination with other regulators.

We also analyzed the RBP network derived from ENCODE eCLIP data and found key regulators associated with cancer. For example, the ENCODE eCLIP experiment has profiled many SUB1 peaks on the 3'UTR regions of genes, and we find that the predicted targets of the RBP SUB1 were significantly up-regulated in many cancer types (Fig. 3C). As a RBP, SUB1 has not been associated with cancer before. We thus validated this new association in liver cancer. After knocking down SUB1 in HepG2 cells, its predicted targets are also down-regulated relative to other genes (Fig. 3D). In addition, we found that the decay rate of SUB1 target genes are significantly shorter than non-targets (Fig. 3C). These results indicate that SUB1 may bind to 3'UTR regions to stabilize transcripts. Moreover, we found that the up-regulation of SUB1 target genes is correlated with a poorer patient survival in other cancer types such as lung cancer (Fig. 4).

We further present the overall regulatory network by systematically arranging it into a hierarchy. TFs are placed into different levels where TFs on the top tend to regulate the expression of other TFs and the ones at the bottom ones are in turn more regulated by others \{cite 25880651}. A final hierarchical network structure is shown in Fig 4. We find that the top layer TFs are not only enriched in cancer associated genes but also more significantly drive tumor-to-normal gene differential expressions.

Extensive rewiring events in regulatory network

For the top-tier cell types with numerous TF ChIP-seq experiments, we constructed cell-type-specific regulatory networks relating to specific cancers and compared them with networks built from their paired normal cell types. We proposed the concept of composite normal by reconciling multiple related normal cell types as shown in figure 5. The pairings -- relating cancerous cell lines to specific tumors and then matching them to normal cell types -- are approximate in nature. However, many of such pairings have been widely used in literature before (see supplementary file). Furthermore, with the enrichment of functional characterization assays in ENCODE, they provide us the first opportunity to directly understand the regulatory alterations in cancer by looking at specific network changes that are "rewired" in the process of oncogenesis.

In "Tumor-normal pairs", we measured the signed, fractional number of edges changing, the rewiring index, to study how the targets of each common TF changed (i.e., rewired) over the course of oncogenic transformation. We first ranked TFs according this index (Fig. 5 A). In leukemia, well-known oncogenes such as MYC and NRF1 were among the top edge gainers, while the well-known tumor suppressor IKZF1 is the most significant edge loser (Fig 5A). Mutations in this latter factor serve as a hallmark of various forms of high-risk leukemia \cite{26202931, 26713593, 26069293 }. Interestingly, IKZF1 loss has been found to be associated with well-known BCR-ABL fusion transcript, which is present in K562, and usually confers poor clinical outcome \cite{26069293}. In contrast, several ubiquitously distributed TFs retain their regulatory linkages (Fig 5A). We observed a similar trend in TFs using a distal, proximal, and combined network (see details in supplementary file). The trend was consistent across highly rewired TFs such as BHLHE40, JUND, and MYC in lung, liver, and breast cancers (Fig 5).

In addition to the simple direct TF to gene connection-based model, we also measured rewiring using more complex gene community model. The targets within the TF regulatory network were characterized by heterogeneous network modules (so called "gene communities"), which usually come from multiple biologically relevant genes. Instead of directly measuring the TF's target changes for each gene, we determined the change in gene communities via a mixed-membership model. This enabled us to evaluate each TF's overall association changes to these gene communities in tumor and normal cells. Similar rewiring patterns were observed using this model (Fig 5A).

We then tested whether the gain or loss events from the normal to tumor transition will result in a network that is more similar or different from those in stem cells like H1-hESC. Interestingly, we find that the gainer group tends to rewire away from the stem cell's regulatory network while the loser groups are more likely to rewire toward the stem cell.

We also find that the majority of rewiring events were associated with noticeable gene expression and chromatin status changes, but not necessarily with variant-induced motif loss or gain events (Fig. 5A). This is consistent with previous discoveries that most non-coding risk variants are not well-explained by the current model \cite{25363779}. For example, JUND is a top gainer in CLL. The majority of its gained targets in tumor cell lines demonstrate higher gene expression, stronger active and weaker repressive histone modification mark signals, yet few of its binding sites are mutated. We found a similar trend for the rewiring events associated with JUND in liver cancer. Related to this, we can formulate the cell-type-specific networks to cell-type-specific hierarchies, as shown in figure 3. Specifically, in blood cancer the more mutationally burdened TFs actually sit at the bottom of the hierarchy, whereas the TFs that are more associated with driving cancer gene expression tend to be at the top

Deleted: . [[

Formatted: Not Highlight

Deleted:]]

Deleted: the mutational burdening of targets

Deleted: XXX

Deleted: that

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Summarizing the analysis described above, the EN-CODEC resource consists of numerous annotation summarized in figure 6 : (1) a BMR model with matching procedure for relevant functional genomics data and a list of regions with higher-than-expected mutations in a diverse selection of different cancers, (2) accurate and refined enhancers and promotors by integrating tens of different functional assays, including STAR-seq, and their comparison with those in ENCODE encyclopedia; (3) enhancer-target-gene linkages and extended gene neighborhoods, based integrating experimentally determined linkages from Hi-C and detailed histone mark and expression correlation, (4) tumor-normal differential expression, chromatin, and regulatory changes, (5) TF regulatory networks, both overall and cell type specific; (6) TFs' position in the network hierarchy and their rewiring status; (7) an analogous but less annotated network for RBPs.

Deleted: and

Deleted: in various formats

Collectively, these resources allow us to prioritize key features as being associated with oncogenesis. The workflow in Fig. 6A describes this prioritization scheme in a systematic fashion. We first search for key regulators that are frequently rewired, located in network hubs or at top of the network hierarchy, or significantly driving expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory changes in terms of expression levels, TF binding, and chromatin status, or are highly mutated in tumors. Finally, on a nucleotide level, we can pinpoint impactful SNVs for small-scale functional characterization by their ability to disrupt or create specific binding sites, or which occur in positions under strong purifying selection.

Using this framework, we subject a number of key regulators, such as MYC and SUB1, to knockdown experiments to validate their regulatory effects in particular cancer contexts (Fig 3D), as we described above. Next here, we also identified several candidate enhancers in noncoding regions, associated with breast cancer, and validated their ability to influence transcription using luciferase assays in MCF7. We selected key SNVs, based on significantly recurrent mutations in breast cancer cohorts, within these enhancers that are important for controlling gene expression. Of the eight motif-disrupting SNVs that we tested, six showed consistent up- or down-regulation relative to the wild type in multiple biological replicates. One particularly interesting example, illustrating the unique value of ENCODE data integration, is in the intronic region of CDH26 in chromosome 20 (Fig. 6C). Both histone modification and chromatin accessibility (DNase-seq) signals indicated an active regulatory role in MCF7, which was further confirmed as an enhancer by both CASPER and ESCAPE (STARR-seq) (Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain (TAD) and validated a regulatory linkage to the downstream breast-cancer-associated gene SYCP2 \cite{26334652, 24662924}. We observed massive binding events from TFs in this region in MCF-7. Motif analysis predicts that the particular mutations found in the cohorts can significantly disrupt the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6D). Luciferase assays demonstrate that this mutation introduces a 3.6-fold reduction in expression relative to wild type expression levels, indicating a strong repressive effect on this enhancer's functionality.

Deleted: as we described above,

Deleted:).

Conclusion

This study highlights the value of our companion to the main ENCODE encyclopedia as a resource for cancer research. First, we show that, by integrating many different types of assays, we can build accurate BMR model for a wide range of cancer and improve the quality and quantity of annotations to look for regions with higher-than-expected mutations. We can also build extensive regulatory networks of various forms from thousands of ChIP-seq and eCLIP experiments to direct study the regulatory alteration during the transformation to cancer and pinpoint key regulators that are involved in cancer progression. Finally, we show how we can leverage the companion resource to provide a prioritization scheme to pinpoint key features for small-scale follow-up.

Our EN-CODEC resource consists two aspects of resources – generalized annotations, such as BMR model and merge networks and hierarchies for pan-cancer type of studies and cancer specific annotations drives from pairing the top-tier cell lines to particular cancer types. We did realize that the representative tumor and normal cell types and their pairings are used here are rough. However, some pairings have already been widely used in other literatures. Besides, cancer is such a heterogeneous disease that even the tumor cells from one patient usually shows distinct molecular, morphological, and genetic profiles \cite{24048065}. It is difficult to obtain a "perfect" match even from data of real tumor and normal tissues. The richness of the ENCODE functional characterization assays does provide us a unprecedented opportunity to systematically study cancer genomes from various aspects.

This study underscores the value of large-scale data integration, and we note that expanding this approach (either by integrating additional data types and/or using tumor mutation and expression data on a larger scale) is straightforward. We also anticipate that an additional step would be to carry out many of the ENCODE assays on specific tissues and tumor samples. For example, larger number of genomic features from matched cell types could result in better BMR estimation; more advanced functional characterization assays will generate compact and accurate annotation sets for larger statistical power in burden analysis; more ChIP-seq/eCLIP experiments would provide more detailed regulatory networks to understand regulatory alterations during cancer progression. In additional, larger cohorts of expression and mutation profiles from many cancer types to discover novel key features for cancer. Though volume of material needed for such analyses may present challenges, we show that such a framework is technically feasible and provides further opportunities for the future.

Deleted: on a large scale

Deleted: a very

Deleted: in various

Deleted: types. We also demonstrated how to improve the statistical power for burden analysis and functional interpretation of the detected mutational hotspots by confining

Deleted: annotation

Deleted: and

Deleted: and improving their linkage to genes

Deleted: Second, w

Deleted: are able to

Deleted: , which are much more accurate than those from imputed binding sites \cite{25409825}. These networks can be directly combined with the expression profiles of various cancer types

Deleted:

Deleted: prioritize

Deleted: Specifically, we also built up cell type specific networks for the first time [[JZ2MG ???]] in the top-tier cell lines and relate them to corresponding normal ones to direct study the regulatory alteration during the transformation to cancer. Then we demonstrate how such comparisons can illuminate potential regulatory changes in cancer (e.g. key rewiring TFs).

Deleted: regulatory elements and SNVs

Deleted: did notice

Deleted: very

Deleted: Finally, we show how we can leverage the companion resource to provide a prioritization scheme to pinpoint key regulatory elements and SNVs for small-scale follow-up.

Deleted: [[more]]